Title	Variable selection in the high-dimensional continuous generalized linear model with current status data
Author(s)	Tian, G; Wang, M; Song, L
Citation	Journal of Applied Statistics, 2014, v. 41 n. 3, p. 467-483
Issued Date	2014
URL	http://hdl.handle.net/10722/203435
Rights	This is an electronic version of an article published in Journal of Applied Statistics, 2014, v. 41 n. 3, p. 467-483. The article is available online at: http://www.tandfonline.com/doi/abs/10.1080/02664763.2013.840271

Variable Selection in the High-Dimensional Continuous Generalized Linear Model with Current Status Data

Guo-Liang Tian

Department of Statistics and Actuarial Science, The University of Hong Kong,

Pokfulam Road, Hong Kong, P. R. China

email: gltian@hku.hk

Mingqiu Wang*

School of Mathematical Sciences, Dalian University of Technology,

Dalian, Liaoning 116023, P. R. China

School of Mathematical Sciences, Qufu Normal University,

Shandong, 273165, P.R.China

**Corresponding author's email*: wmq0829@gmail.com

Lixin Song

School of Mathematical Sciences, Dalian University of Technology,

Dalian, Liaoning 116023, P. R. China

email: lxsong@dlut.edu.cn

Summary. In survival studies, current status data are frequently encountered when some individuals in a study are not successively observed. This paper considers the problem of simultaneous variable selection and parameter estimation in the high-dimensional continuous generalized linear model with current status data. We apply the penalized likelihood procedure with the SCAD penalty to select significant variables and estimate the corresponding regression coefficients. With a proper choice of tuning parameters, the resulting estimator is shown to be a root n/p_n-consistent estimator under some mild conditions. In addition, we show that the resulting estimator has the same asymptotic distribution as the estimator ob-

tained when the true model is known. The finite sample behavior of the proposed estimator is evaluated through simulation studies and a real example.

KEY WORDS: Current status data; Generalized linear model; Oracle property; SCAD penalty; Variable selection.

1 Introduction

In survival studies, the random survival time of interest is often too expensive or even impossible to observe the exact time. However, the current status at a random inspection time is much more practical. Examples of current status data include clinical study of tumor occurrence (Gart *et al.*, 1986), HIV transmission among sexual partners (Jewell and Shiboski, 1990), demographic study of age at weaning (Grummer-Strawn, 1993), and so on. Such data structure is called case I interval-censored data (which is a type of interval-censored data) or current status data. The analysis of current status data arising frequently in medical research has recently attracted a great amount of attention (Huang, 1996; Xue *et al.*, 2004; Lam and Xue, 2005; Ma, 2009; Lin and Wang, 2010; Wang and Lin, 2011).

Notice the difference between current status data and usual right censoring data. They are quite different in terms of their structures and the information contained. In particular, their censoring mechanisms are different. For the current status data, the survival times of interest are only known to be either left-censored or right-censored. In other words, current status data mean that each observed interval for the survival variable includes either zero or infinity. Compared to right-censored data, current status data contain much less information about the survival variable of interest. Therefore, most of the inference procedures developed for right-censored data cannot be easily/directly applied to current status data.

Variable selection is an important topic in contemporary statistics. Much progress has been made in exploring the variable selection and statistical properties for high dimensional

data. Various penalized approaches have been successively proposed. Examples include the bridge penalty (Frank and Friedman, 1993), the *least absolute shrinkage and selection operator* (Lasso, Tibshirani, 1996), the *smoothly clipped absolute deviation* (SCAD) penalty (Fan and Li, 2001), the Dantzig selector (Candes and Tao, 2007), and the *minimum concave penalty* (MCP, Zhang, 2010). There are a large number of researches about variable selection for high dimensional uncensored data. The readers can refer to Fan and Peng (2004), Huang *et al.* (2008), Fan and Lv (2008), Bradic *et al.* (2010), Wang *et al.* (2010), Wang *et al.* (2011), and references therein. In the last decades, much work has been done on the variable selection for right-censored data. Huang *et al.* (2006) considered the variable selection in the accelerated failure time model with diverging dimensions. Huang and Ma (2010) studied the variable selection in the accelerated failure time model via the bridge penalty. Ma and Du (2011) studied the variable selection in the partially linear model with high-dimensional covariates. However, all these results cannot be directly generalized to the current status data due to the aforementioned differences between the current status data and right-censored data.

Up to now, it seems that there is no systematic theoretical investigation of simultaneous variable selection and coefficients estimation in the continuous generalized linear model with current status data. The main purpose of our paper is to fill in this gap. In this paper, we study some asymptotic properties of estimators in the high dimensional generalized linear model with current status data when the number of covariates diverges with the sample size. Here, we assume the response variable is continuous. In order to achieve simultaneous variable selection and parameters estimation, we define a penalized log-likelihood function with the SCAD penalty. With a proper choice of regularization parameters, the resulting estimator is shown to be a root n/p_n-consistent estimator under some mild conditions. Furthermore, we show that the resulting estimator has the same asymptotic distribution as the estimator obtained when the true model is known.

The rest of the paper is organized as follows. Section 2 presents the continuous generalized linear model with current status data and the penalized log-likelihood function. Asymptotic properties of the penalized likelihood estimator are provided in Section 3. Section 4 discusses the computation of the estimates and the choice of tuning parameters. In addition, two simulation studies are conducted and a real dataset is analyzed to illustrate the finite sample performance of the proposed method. A discussion is presented in Section 5. All technical proofs are given in the Appendix.

2 Model and penalized likelihood

2.1 Continuous generalized linear model with current status data

Consider the continuous generalized linear regression model

$$Y = g(\boldsymbol{\beta}_n^\top \boldsymbol{X}) + \varepsilon, \tag{2.1}$$

where Y is a continuous response variable, the inverse of $g(\cdot)$ is a known smooth link function, $\boldsymbol{\beta}_n$ is an unknown $p_n \times 1$ vector of regression coefficients, $\boldsymbol{X}$ is a $p_n \times 1$ random vector of predictors, and ε is a random error with mean 0. Here the subscript n is indicated that variables are allowed to diverge with n. Suppose that ε has a cumulative distribution function $F(\cdot)$ and a corresponding density function $f(\cdot)$, where $f(\cdot)$ is assumed to have a finite second derivative. In addition, we assume that $g(\cdot)$ has a finite third derivative.

In this paper, we consider the model (2.1) to fit case I interval-censored data (i.e., current status data). In other words, the response variable of interest Y cannot be observed directly, but $\delta = I(Y \leqslant Z)$ can be observed, where $I(\cdot)$ denotes the indicator function, Z is a censoring random variable with density $h(z)$, and Z is independent of $\boldsymbol{X}$. We further assume that the density $\varphi(\boldsymbol{x})$ of the covariate vector $\boldsymbol{X}$ is known. In addition, assume that ε is independent of $(\boldsymbol{X}, Z)$. Let the observable random vector be $\mathbf{W} = (\delta, \boldsymbol{X}, Z)$. The density of $\mathbf{W}$ is given by

$$[F(z - g(\boldsymbol{\beta}_n^\top \boldsymbol{x}))]^\delta [1 - F(z - g(\boldsymbol{\beta}_n^\top \boldsymbol{x}))]^{1-\delta} \varphi(\boldsymbol{x}) h(z).$$

Since $\varphi(\boldsymbol{x})$ and $h(z)$ do not involve the unknown parameter vector $\boldsymbol{\beta}_n$, we can treat them as constants in the estimation of $\boldsymbol{\beta}_n$. So the log-likelihood function is proportional to

$$\delta \log[F(Z - g(\boldsymbol{\beta}_n^\top \boldsymbol{X}))] + (1-\delta)\log[1 - F(Z - g(\boldsymbol{\beta}_n^\top \boldsymbol{X}))].$$

Let $\mathbf{W}_1, \ldots, \mathbf{W}_n$ be an independent and identically distributed (i.i.d.) sample distributed as $\mathbf{W}$, where $\mathbf{W}_i = (\delta_i, \mathbf{X}_i, Z_i)$. The log-likelihood function for the observed sample is $\ell_n(\boldsymbol{\beta}_n) = \sum_{i=1}^n \ell_{ni}(\boldsymbol{\beta}_n)$, where

$$\ell_{ni}(\boldsymbol{\beta}_n) = \delta_i \log[F(Z_i - g(\boldsymbol{\beta}_n^\top \mathbf{X}_i))] + (1-\delta_i)\log[1 - F(Z_i - g(\boldsymbol{\beta}_n^\top \mathbf{X}_i))].$$

2.2 Variable selection methods

In recent literature, there are several versions about the penalty function. The bridge penalty was originally proposed by Frank and Friedman (1993) corresponding to the L_q-penalty $p_\lambda(|\theta|) = \lambda|\theta|^q$. Tibshirani (1996) studied the Lasso penalty for more details. Knight and Fu (2000) investigated the L_q-penalty with $q < 1$. Fan and Li (2001) advocated the SCAD penalty, which is defined by

$$p_\lambda(|\theta|) = \begin{cases} \lambda|\theta|, & \text{if } 0 \leqslant |\theta| < \lambda, \\ \dfrac{(a^2-1)\lambda^2 - (|\theta| - a\lambda)^2}{2(a-1)}, & \text{if } \lambda \leqslant |\theta| < a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & \text{if } |\theta| \geqslant a\lambda, \end{cases}$$

where $a > 2$ and $\lambda > 0$ are the tuning parameters. The SCAD penalty is continuous and differentiable on $(-\infty, 0) \cup (0, \infty)$, but not differentiable at 0. Its derivative vanishes outside $[-a\lambda, a\lambda]$. Hence, the SCAD penalty can produce continuity, sparsity and unbiasedness estimator for large coefficients. More details can be found in Fan and Li (2001). Zou (2006) proposed the adaptive Lasso with form $p_\lambda(|\theta|) = \lambda w|\theta|$, where w is a weight. Zhang (2010) gave the *minimax concave penalty* (MCP) which performs as well as the SCAD penalty and the adaptive Lasso. The MCP is defined as

$$p(\theta; \lambda, \gamma) = \lambda \int_0^{|\theta|} (1 - x/(\gamma\lambda))_+ \, \mathrm{d}x.$$

In this paper, to emphasize the dependency of λ on n, we denote λ by λ_n. In addition, as suggested by Fan and Li (2001), we fix $a = 3.7$.

2.3 Penalized likelihood function

Consider the penalized likelihood function for estimating $\boldsymbol{\beta}_n$ as follows

$$Q_n(\boldsymbol{\beta}_n) = \ell_n(\boldsymbol{\beta}_n) - n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|), \tag{2.2}$$

where the function $p_{\lambda_n}(\cdot)$ is the SCAD penalty.

Let the true parameter value be $\boldsymbol{\beta}_{n0}$, but for simplicity, we will write it as $\boldsymbol{\beta}_0$. In the sparse model, some components of covariates are trivial and the corresponding coefficients are zero. For convenience of notation, let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^\top, \boldsymbol{\beta}_{20}^\top)^\top$, where $\boldsymbol{\beta}_{10}^\top = (\beta_{01}, \ldots, \beta_{0k_n})$ is a $k_n \times 1$ vector and $\boldsymbol{\beta}_{20}^\top = (0, \ldots, 0)$ is an $m_n \times 1$ vector. Here k_n is the number of nonzero coefficients and $m_n = p_n - k_n$ is the number of trivial covariates. Similarly, we can partition the population vector of covariates $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$ and the corresponding sample $\mathbf{X}_i = (\mathbf{X}_{1i}^\top, \mathbf{X}_{2i}^\top)^\top$, where $\mathbf{X}_{1i} = (X_{i1}, \ldots, X_{ik_n})^\top$ and $\mathbf{X}_{2i} = (X_{i(k_n+1)}, \ldots, X_{ip_n})^\top$.

3 Asymptotic properties of penalized likelihood estimator

In this section, we establish several theoretical properties of the penalized likelihood estimator when the number of predictors increases with the sample size. First, we define some notations.

Let $\xi(\boldsymbol{\beta}_n, \mathbf{W}_i) = Z_i - g(\boldsymbol{\beta}_n^\top \mathbf{X}_i)$,

$$\begin{aligned} D(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i)) &= \frac{\partial \ell_{ni}(\boldsymbol{\beta}_n)}{\partial \xi(\boldsymbol{\beta}_n, \mathbf{W}_i)} \\ &= \frac{\delta_i f(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i))}{F(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i))} - \frac{(1-\delta_i) f(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i))}{1 - F(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i))}, \quad i = 1, \ldots, n. \end{aligned}$$

So

$$\begin{aligned}
\frac{\partial \ell_{ni}(\boldsymbol{\beta}_n)}{\partial \boldsymbol{\beta}_n} &= \frac{\partial \ell_{ni}(\boldsymbol{\beta}_n)}{\partial \xi(\boldsymbol{\beta}_n, \mathbf{W}_i)} \frac{\partial \xi(\boldsymbol{\beta}_n, \mathbf{W}_i)}{\partial \boldsymbol{\beta}_n} = D(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i))[-g'(\boldsymbol{\beta}_n^\top \mathbf{X}_i)\mathbf{X}_i], \\
\frac{\partial^2 \ell_{ni}(\boldsymbol{\beta}_n)}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^\top} &= D'(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i))[g'(\boldsymbol{\beta}_n^\top \mathbf{X}_i)]^2 \mathbf{X}_i \mathbf{X}_i^\top - D(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i)) g''(\boldsymbol{\beta}_n^\top \mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^\top \\
&\triangleq d_{n1}(\boldsymbol{\beta}_n, \mathbf{W}_i) \mathbf{X}_i \mathbf{X}_i^\top, \\
\frac{\partial^3 \ell_{ni}(\boldsymbol{\beta}_n)}{\partial \beta_{nj} \partial \beta_{nk} \partial \beta_{nl}} &= \Big\{ - D''(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i))[g'(\boldsymbol{\beta}_n^\top \mathbf{X}_i)]^3 + 3D'(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i)) g''(\boldsymbol{\beta}_n^\top \mathbf{X}_i) g'(\boldsymbol{\beta}_n^\top \mathbf{X}_i) \\
&\quad - D(\xi(\boldsymbol{\beta}_n, \mathbf{W}_i)) g^{(3)}(\boldsymbol{\beta}_n^\top \mathbf{X}_i) \Big\} X_{ij} X_{ik} X_{il} \\
&\triangleq d_{n2}(\boldsymbol{\beta}_n, \mathbf{W}_i) X_{ij} X_{ik} X_{il}.
\end{aligned}$$

Let $P_{\boldsymbol{\beta}_n}$ be the distribution function of $\mathbf{W}$ and E_0 be the expectation with respect to $P_{\boldsymbol{\beta}_0}$. For simplicity, the main assumptions required for our results are presented as follows.

(A1) $E_0[D(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1)) g'(\boldsymbol{\beta}_0^\top \mathbf{X}_1) \mathbf{X}_1] = \mathbf{0}$.

(A2) For $j, k = 1, \ldots, p_n$,

$$E_0\{D^2(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1))[g'(\boldsymbol{\beta}_0^\top \mathbf{X}_1)]^2 X_{1j} X_{1k}\} = -E_0[d_{n1}(\boldsymbol{\beta}_0, \mathbf{W}_1) X_{1j} X_{1k}].$$

(A3) The Fisher information matrix

$$\begin{aligned}
\mathrm{I}_n(\boldsymbol{\beta}_0) &= E_0\Big\{[-D(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1)) g'(\boldsymbol{\beta}_0^\top \mathbf{X}_1)\mathbf{X}_1][-D(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1)) g'(\boldsymbol{\beta}_0^\top \mathbf{X}_1)\mathbf{X}_1]^\top\Big\} \\
&= E_0\Big\{D^2(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1))[g'(\boldsymbol{\beta}_0^\top \mathbf{X}_1)]^2 \mathbf{X}_1 \mathbf{X}_1^\top\Big\}.
\end{aligned}$$

Let the smallest and largest eigenvalues of $\mathrm{I}_n(\boldsymbol{\beta}_0)$ be $\lambda_{\min}\{\mathrm{I}_n(\boldsymbol{\beta}_0)\}$ and $\lambda_{\max}\{\mathrm{I}_n(\boldsymbol{\beta}_0)\}$, which satisfy

$$0 < M_1 \leqslant \lambda_{\min}\{\mathrm{I}_n(\boldsymbol{\beta}_0)\} \leqslant \lambda_{\max}\{\mathrm{I}_n(\boldsymbol{\beta}_0)\} \leqslant M_2 < \infty,$$

where M_1 and M_2 are given constants.

(A4) There exist constants $0 < M_3, M_4, M_5 < \infty$ such that

$$\max_{1 \leqslant j \leqslant p_n} |X_{1j}| \leqslant M_3, \quad E_0[d_{n1}(\boldsymbol{\beta}_0, \mathbf{W}_1)]^2 \leqslant M_4$$

and

$$E_0[D(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1))g'(\boldsymbol{\beta}_0^\top \mathbf{X}_1)]^4 \leqslant M_5.$$

(A5) There is a large enough open subset $\mathbb{S}_n$ that contains the true parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_n}$ and a function $H(\mathbf{W}_i)$ such that, for all $\boldsymbol{\beta}_n \in \mathbb{S}_n$, $|d_{n2}(\boldsymbol{\beta}_n, \mathbf{W}_i)| \leqslant H(\mathbf{W}_i), i = 1, \ldots, n$. In addition, there exists a constant M_6 such that $E_0[H^2(\mathbf{W}_1)] \leqslant M_6$.

(A6) ρ_{n1} and ρ_{n2} are bounded away from zero, where ρ_{n1} and ρ_{n2} are the smallest and largest eigenvalue of $E_0(H(\mathbf{W}_1)\mathbf{X}_1\mathbf{X}_1^\top)$, respectively.

These conditions are needed to obtain the asymptotic results in the theorems below. Condition (A1) is easy to check. Conditions (A2)–(A5) are similar to regularity conditions that guarantee asymptotic properties of the maximum likelihood estimators without censoring (Fan and Peng, 2004). Here we impose them to facilitate the technical proof. For example, we could impose some more detailed restrictions on the parameter space and functions f and g instead of condition (A4). The form of $\mathrm{I}_n(\boldsymbol{\beta}_0)$ in condition (A3) is similar to that in Xue *et al.* (2004).

Theorem 3.1 (CONSISTENCY). Suppose $n\lambda_n^2 = O(1)$ and $p_n^3/n \to 0$. Then under conditions (A1)–(A6), there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ of $Q_n(\boldsymbol{\beta}_n)$ such that

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_P(\sqrt{p_n/n}\,). \qquad \square$$

Remark 3.1 Theorem 3.1 shows that we can obtain the consistent estimator even when the data are censored. Under some regular conditions, the convergence rate is optimal for the case of diverging number of parameters. $\square$

Theorem 3.2 (ORACLE PROPERTY). Suppose that $p_n^3/n \to 0$, $\min_{1\leqslant j\leqslant k_n} |\beta_{0j}|/\lambda_n \to \infty$, and $\sqrt{p_n/n}/\lambda_n \to 0$. If conditions (A1)–(A6) are satisfied, then the local maximizer $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^\top, \hat{\boldsymbol{\beta}}_{2n}^\top)^\top$ in Theorem 3.1 satisfies

(1) Sparsity:

$$\Pr(\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}) \to 1 \quad \text{as} \quad n \to \infty.$$

(2) Asymptotic normality:

$$\sqrt{n}\boldsymbol{\alpha}^\top \mathrm{I}_n^{\frac{1}{2}}(\boldsymbol{\beta}_{10})(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \xrightarrow{D} N(0,1),$$

where $\boldsymbol{\alpha}$ is an arbitrary $k_n \times 1$ vector with $\|\boldsymbol{\alpha}\| = 1$. □

Remark 3.2 Theorem 3.2 indicates that under certain conditions on the tuning parameter and model, the estimator enjoys the oracle property. Note that the model studied in this paper has a specific density, so we only need the condition $p_n^3/n \to 0$ after some detailed calculations, which is weaker than the conditions $p_n^4/n \to 0$ and $p_n^5/n \to 0$ (Fan and Peng, 2004). □

4 Numerical examples

In this section, we first present an algorithm to conduct the variable selection. Next, several simulation experiments are carried out to assess the finite sample performance of the proposed method. Finally, a real data set is used to the illustration.

4.1 Computational algorithm

4.1.1 Local quadratic approximation and standard errors

Fan and Li (2001) proposed the *local quadratic approximation* (LQA) algorithm to attack the singularity of the SCAD penalty at the origin. In this paper, we apply the LQA algorithm to obtain the regression coefficient estimate in the continuous generalized linear model for

the current status data. Suppose that there is an initial value $\boldsymbol{\beta}_n^{(0)}$ that is very close to the maximizer of (2.2). If $\beta_{nj}^{(0)}$ is very close to 0, then we set $\beta_{nj}^{(0)} = 0$. Otherwise, the penalty function $p_{\lambda_n}(|\beta_{nj}|)$ can be locally approximated by the following function

$$p_{\lambda_n}(|\beta_{nj}|) \approx p_{\lambda_n}(|\beta_{nj}^{(0)}|) + \frac{1}{2}\frac{p'_{\lambda_n}(|\beta_{nj}^{(0)}|)}{|\beta_{nj}^{(0)}|}(\beta_{nj}^2 - \beta_{nj}^{(0)2}), \quad \text{for } \beta_{nj} \approx \beta_{nj}^{(0)}. \tag{4.1}$$

Replacing the penalty function in (2.2) by (4.1), we can use the Newton–Raphson algorithm to find the maximizer of (2.2). In fact, for the initial value $\boldsymbol{\beta}_n^{(0)}$, the log-likelihood function $\ell_n(\boldsymbol{\beta}_n)$ can be locally approximated by

$$\ell_n(\boldsymbol{\beta}_n^{(0)}) + \left[\frac{\partial \ell_n(\boldsymbol{\beta}_n^{(0)})}{\partial \boldsymbol{\beta}_n}\right]^\top (\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^{(0)}) + \frac{1}{2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^{(0)})^\top \left[\frac{\partial^2 \ell_n(\boldsymbol{\beta}_n^{(0)})}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^\top}\right](\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^{(0)}). \tag{4.2}$$

Therefore, by combining (4.1) with (4.2), we can see that the maximization of (2.2) is equivalent to the maximization of the following expression

$$\begin{aligned}&\left[\frac{\partial \ell_n(\boldsymbol{\beta}_n^{(0)})}{\partial \boldsymbol{\beta}_n}\right]^\top (\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^{(0)}) + \frac{1}{2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^{(0)})^\top \left[\frac{\partial^2 \ell_n(\boldsymbol{\beta}_n^{(0)})}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^\top}\right](\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^{(0)}) \\ &- \frac{1}{2} n \boldsymbol{\beta}_n^\top \Sigma_{\lambda_n}(\boldsymbol{\beta}_n^{(0)}) \boldsymbol{\beta}_n,\end{aligned} \tag{4.3}$$

where

$$\Sigma_{\lambda_n}(\boldsymbol{\beta}_n^{(0)}) = \text{diag}\left\{\frac{p'_{\lambda_n}(|\beta_{n1}^{(0)}|)}{|\beta_{n1}^{(0)}|}, \ldots, \frac{p'_{\lambda_n}(|\beta_{np_n}^{(0)}|)}{|\beta_{np_n}^{(0)}|}\right\}.$$

Accordingly, the quadratic maximization of (4.3) leads to the following iteration:

$$\boldsymbol{\beta}_n^{(1)} = \boldsymbol{\beta}_n^{(0)} - \left[\frac{\partial^2 \ell_n(\boldsymbol{\beta}_n^{(0)})}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^\top} - n\Sigma_{\lambda_n}(\boldsymbol{\beta}_n^{(0)})\right]^{-1} \left[\frac{\partial \ell_n(\boldsymbol{\beta}_n^{(0)})}{\partial \boldsymbol{\beta}_n} - n\Sigma_{\lambda_n}(\boldsymbol{\beta}_n^{(0)})\boldsymbol{\beta}_n^{(0)}\right]. \tag{4.4}$$

The estimator of parameter $\boldsymbol{\beta}_n$ can be obtained according to the following algorithm:

Step 1. Let the initial value $\boldsymbol{\beta}_n^{(0)}$ equal the ordinary maximum likelihood estimate (without penalty). If $|\beta_{nj}^{(0)}| < \tau$ (τ is a pre-specified constant and equals to 10^{-4} in our simulations and application to real data), then set $\beta_{nj}^{(0)} = 0$.

Step 2. Given the current value $\boldsymbol{\beta}_n^{(k)} = \boldsymbol{\beta}_n^{(0)}$, we can obtain $\boldsymbol{\beta}_n^{(k+1)}$ by the formula (4.4).

Step 3. Repeat Step 2 until $\max_{1\leqslant j\leqslant p_n}|\beta_{nj}^{(k+1)}-\beta_{nj}^{(k)}|\leqslant\tau$.

Using the similar techniques in Fan and Peng (2004), the covariance matrix of $\hat{\boldsymbol{\beta}}_{1n}$ (the nonzero components of $\hat{\boldsymbol{\beta}}_n$), can be approximated by the following sandwich formula:

$$\left[\frac{\partial^2\ell_n(\hat{\boldsymbol{\beta}}_{1n})}{\partial\boldsymbol{\beta}_{1n}\partial\boldsymbol{\beta}_{1n}^{\mathrm{T}}}-n\Sigma_{\lambda_n}(\hat{\boldsymbol{\beta}}_{1n})\right]^{-1}\widehat{\mathrm{cov}}\left(\frac{\partial\ell_n(\hat{\boldsymbol{\beta}}_{1n})}{\partial\boldsymbol{\beta}_{1n}}\right)\left[\frac{\partial^2\ell_n(\hat{\boldsymbol{\beta}}_{1n})}{\partial\boldsymbol{\beta}_{1n}\partial\boldsymbol{\beta}_{1n}^{\mathrm{T}}}-n\Sigma_{\lambda_n}(\hat{\boldsymbol{\beta}}_{1n})\right]^{-1},$$

where $\widehat{\mathrm{cov}}(\partial\ell_n(\hat{\boldsymbol{\beta}}_{1n})/\partial\boldsymbol{\beta}_{1n})$ is the covariance matrix of $\partial\ell_n(\boldsymbol{\beta}_{1n})/\partial\boldsymbol{\beta}_{1n}$ evaluated at $\boldsymbol{\beta}_{1n}=\hat{\boldsymbol{\beta}}_{1n}$.

4.1.2 Choice of the tuning parameter

It is very critical to choose a proper tuning parameter λ_n since it determines the sparsity of the selected model. An optimal tuning parameter can result in a parsimonious model with good prediction performance. Wang *et al.* (2007, 2009) showed that Bayesian information criterion (BIC) is consistent in model selection. We employ the BIC-type criterion to choose the tuning parameter. For a given λ_n, we can obtain an estimate $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. Let d_{λ_n} be the number of nonzero components of $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. The BIC-type criterion is defined by

$$\mathrm{BIC}(\lambda_n)=-2\ell_n(\widehat{\boldsymbol{\beta}}_{\lambda_n})+d_{\lambda_n}\times\log n.$$

4.2 Simulation studies

In this subsection, we evaluate the performance of the proposed method through two simulation examples. To measure the estimation accuracy of the estimator, we use the average *mean squared errors* (MSE) $E\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|^2$. The variable selection performance is assessed by (C, IC, Correctly fitted, Overfitted), where "C" denotes the average number of zero coefficients correctly set to zero, "IC" is the average number of nonzero coefficients incorrectly set to zero, "Correctly fitted" represents the proportion of times that the correct model is selected, and "Overfitted" is the proportion of including all significant variables and some noise variables. We compare the performance of the SCAD penalty with the Lasso, the

adaptive Lasso (ALasso) and the Oracle. The oracle estimator is computed by using the true model when the zero coefficients are known. In practice, the oracle estimator cannot be obtained. We only use it as a benchmark for comparison. For each simulation setting, 500 simulated data sets are generated.

EXAMPLE 1. Let n observations be generated from the linear model

$$Y = \boldsymbol{\beta}_n^\top \boldsymbol{X} + \varepsilon,$$

where $\boldsymbol{X} = (X_1, \ldots, X_{p_n})^\top$. The number of parameter is assumed to be $p_n = \lfloor 6n^{1/4} - 5 \rfloor$ and the number of nonzero coefficients is assumed to be $k_n = 3q_n$, where $q_n = \lfloor p_n/7 \rfloor$ and $\lfloor \cdot \rfloor$ denotes the floor function. The true coefficients $\boldsymbol{\beta}_n^\top = (0.8 \cdot \mathbf{1}_{q_n}^\top, \mathbf{1}_{q_n}^\top, 1.5 \cdot \mathbf{1}_{q_n}^\top, \mathbf{0}_{p_n - k_n}^\top)$, where $\mathbf{1}_m$ is an m-vector of ones and $\mathbf{0}_m$ is an m-vector of zeros. X_j $(j = 1, \ldots, p_n)$ are independent standard normal variables. We consider two different error distributions. The first error follows the standard normal distribution and the censoring variable Z is generated from $N(\mu_1, 1)$ for each simulated data set, where μ_1 is chosen such that the corresponding censoring rate is about 25%. The second error has a standard logistic distribution and the censoring variable $Z \sim \text{Logistic}(\mu_2, 1)$, where μ_2 is chosen to obtain the censoring rate 25%. We consider three sample sizes, $n = 100, n = 300$ and $n = 600$.

Table 1 summarizes the average MSE and the corresponding results of variable selection. The numbers in parentheses are standard deviations. From Table 1, it is easy to see that

(1) Overall, both SCAD and adaptive Lasso perform better than the Lasso in terms of both variable selection and MSE. The SCAD outperforms the adaptive Lasso when the sample size is large. When the sample size increases, for the SCAD, the proportion of times of the correctly selected model increases while the MSE decreases. Although the Lasso can produce a sparse model, the proportion of times of the correctly selected model is very low for large sample sizes.

(2) In terms of MSE, there exists a certain discrepancy between SCAD and Oracle for small sample sizes. However, the discrepancy becomes very small when the sample size

increases to infinity. In contrast, although the discrepancy between Lasso and Oracle also decreases when sample sizes increase, the discrepancy is still very significant for large sample sizes. Therefore, we can conclude that the SCAD enjoys the oracle property as the sample size tends to infinity, while the Lasso does not.

(3) For the normal and logistic error distributions, both Lasso and SCAD can identify redundant parameters and reduce the complexity of the model. When the quasi-likelihood method is applied to the continuous generalized linear model, we find that the results for variable selection have no significant difference for the two error distributions.

(4) As suggested by one referee, we show the difference about results of the SCAD by choosing the value of a (denote by SCAD* in Table 1). From Table 1, we can see that the choice of $a = 3.7$ is very reasonable, especially for large sample sizes.

EXAMPLE 2. In this example, we generate n observations from

$$Y = \exp(\boldsymbol{\beta}_n^\top \boldsymbol{X}) + \varepsilon.$$

The true regression coefficients are set to be $\boldsymbol{\beta}_n^\top = (0.4 \cdot \mathbf{1}_{q_n}^\top, 0.5 \cdot \mathbf{1}_{q_n}^\top, 0.75 \cdot \mathbf{1}_{q_n}^\top, \mathbf{0}_{p_n - k_n}^\top)$, while the other parameters are identical to those in Example 1. The simulation results are displayed in Table 2. The numbers in parentheses are the corresponding standard errors. From Table 2, we can obtain a similar conclusion as in Example 1.

4.3 Application to primary biliary cirrhosis data

Consider the primary biliary cirrhosis (PBC) data of the liver collected from January 1974 to May 1984 in Mayo Clinic trial for comparing the drug D-penicillamine (DPCA) with a placebo. The data contain information about the survival time and prognostic factors for 418 patients. Discarding observations with missing values, only 276 observations are available. Variables in this dataset include survival time T_i, right censoring indicator δ_i, and 17 covariates $X_1, \ldots, X_{17}$. All the notations are the same as those of Tibshirani (1997). The

detailed descriptions of this dataset can be found in Fleming and Harrington (1991) and Tibshirani (1997), where the Cox model is employed in their analyses. Here, we treat these data as the current status data and apply the linear model as an illustration. We take the logarithm transformation to T_i and standardize the covariates.

Table 3 gives the estimated coefficients of four methods including the maximum likelihood estimate (MLE), Lasso, adaptive Lasso (ALasso) and SCAD, together with the corresponding standard errors. We also list the results for the Lasso in Tibshirani (1997) for comparison (Lasso(T)). The optimal values of λ_n are 0.053, 0.012 and 0.082 for the Lasso, adaptive Lasso and SCAD, respectively. From Table 3, we find that the SCAD identifies a simpler model with seven important variables, while the Lasso includes more variables. The adaptive Lasso contains ten variables which are included by the Lasso. For the Lasso, We can see that our results are same as that of Tibshirani (1997) except for the variable "Alkaline phosphatase", which is selected by others such as Shows *et al.* (2010).

5 Discussion

When comparing with the right-censored data, the current status data provides less information for analysis, resulting in some challenges in statistical inferences. The existing studies about modeling the current status data mainly focus on the estimation of the regression coefficients. Little work has been done on the variable selection in the setting of current status data. In this paper, we study variable selection about the high-dimensional continuous generalized linear model with current status data. We apply the SCAD penalty to achieve the identification of the sparsity model. Under some regularity conditions, the rate of convergence of the proposed estimator and oracle property are established when the numbers of parameters increase to infinity as the sample size. The effectiveness of the proposed method is verified through simulation studies and a real data set.

We demonstrate the convergence of our algorithm. The data are generated from the model in example 1 (Section 4.2). The sample size is 100, and the error is standard normal distribution. Our experiment showed that the proposed algorithm converged to the right solution. The corresponding computation time in R for the SCAD, adaptive Lasso and Lasso are 0.37, 0.57 and 0.49 s, respectively. The numbers of iterations are 10, 11 and 33, respectively for the SCAD, adaptive Lasso and Lasso.

We have only considered the SCAD penalty. It is not difficult to obtain the variable selection results via the MCP function, because both the SCAD and MCP belong to nonconvex penalty. In addition, how to derive the theoretical properties in the setting of ultrahigh dimensionality is an interesting topic for our future study.

A Appendix

To facilitate the proof of Theorem 3.1, we need the following result.

Lemma A.1 Under conditions (A1), (A2) and (A4), if $p_n^3/n \to 0$, then we have

$$\left\| \frac{1}{n}\nabla^2 \ell_n(\boldsymbol{\beta}_0) + \mathrm{I}_n(\boldsymbol{\beta}_0) \right\| = o_P\left(\frac{1}{\sqrt{p_n}}\right). \qquad \square$$

Proof. For any $\epsilon > 0$, by the Chebyshev's inequality, we have

$$\begin{aligned}
& \Pr\left\{ \left\| \frac{1}{n}\nabla^2 \ell_n(\boldsymbol{\beta}_0) + \mathrm{I}_n(\boldsymbol{\beta}_0) \right\| \geqslant \frac{1}{\sqrt{p_n}}\epsilon \right\} \\
\leqslant & \frac{p_n}{\epsilon^2}\frac{1}{n^2} E_0 \|\nabla^2 \ell_n(\boldsymbol{\beta}_0) + n\mathrm{I}_n(\boldsymbol{\beta}_0)\|^2 \\
= & \frac{p_n}{\epsilon^2}\frac{1}{n^2} \sum_{j,k=1}^{p_n} nE_0 \left[\frac{\partial^2 \ell_{n1}(\boldsymbol{\beta}_0)}{\partial\beta_{nj}\partial\beta_{nk}} + \mathrm{I}_{njk}(\boldsymbol{\beta}_0) \right]^2 \\
\leqslant & \frac{p_n}{\epsilon^2}\frac{1}{n} \sum_{j,k=1}^{p_n} E_0[d_{n1}(\boldsymbol{\beta}_0, \mathbf{W}_1) X_{1j} X_{1k}]^2 \\
\leqslant & \frac{p_n}{n\epsilon^2} p_n^2 M_3^4 M_4 \to 0.
\end{aligned}$$

□

Proof of Theorem 3.1. It suffices to show that for any $\epsilon > 0$, there exists a large constant $C > 0$ such that

$$\Pr\left\{\sup_{\|\mathbf{u}\|=C} Q_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) < Q_n(\boldsymbol{\beta}_0)\right\} \geqslant 1 - \epsilon, \tag{A.1}$$

where $\alpha_n \triangleq \sqrt{p_n/n}$. (A.1) implies that with probability at least $1-\epsilon$, there exists a local maximum in the ball $\{\boldsymbol{\beta}_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leqslant C\}$, where $\mathbf{u}$ is a $p_n \times 1$ scalar vector. That is, there exists a local maximizer such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_P(\sqrt{p_n/n}\,)$.

Noting that $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned}
& Q_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - Q_n(\boldsymbol{\beta}_0) \\
= \;& \ell_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell_n(\boldsymbol{\beta}_0) - n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{0j} + \alpha_n u_j|) + n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{0j}|) \\
\leqslant \;& \Big[\ell_n(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell_n(\boldsymbol{\beta}_0)\Big] + n\sum_{j=1}^{k_n} p_{\lambda_n}(|\beta_{0j}|) \\
\triangleq \;& I_{n1} + I_{n2}.
\end{aligned}$$

First, we consider the term I_{n1}. Applying the third order Taylor expansion, we obtain

$$\begin{aligned}
I_{n1} &= \alpha_n \nabla^\top \ell_n(\boldsymbol{\beta}_0)\mathbf{u} + \frac{1}{2}\alpha_n^2 \mathbf{u}^\top \nabla^2 \ell_n(\boldsymbol{\beta}_0)\mathbf{u} + \frac{1}{6}\alpha_n^3 \nabla^\top \left(\mathbf{u}^\top \nabla^2 \ell_n(\boldsymbol{\beta}_n^*)\mathbf{u}\right)\mathbf{u} \\
&\triangleq I_{n11} + I_{n12} + I_{n13}.
\end{aligned}$$

For the first term I_{n11}, by the conditions (A1) and (A3), we obtain

$$\begin{aligned}
E_0(I_{n11}^2) &= \alpha_n^2 E_0\left\{\sum_{i=1}^{n}[-D(\xi(\boldsymbol{\beta}_0, \mathbf{W}_i))]g'(\boldsymbol{\beta}_0^\top \mathbf{X}_i)\mathbf{X}_i^\top \mathbf{u}\right\}^2 \\
&= n\alpha_n^2 E_0[D^2(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1))g'^2(\boldsymbol{\beta}_0^\top \mathbf{X}_1)(\mathbf{X}_1^\top \mathbf{u})^2] \\
&= n\alpha_n^2 \mathbf{u}^T E_0[D^2(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1))g'^2(\boldsymbol{\beta}_0^\top \mathbf{X}_1)\mathbf{X}_1\mathbf{X}_1^\top]\mathbf{u} \\
&\leqslant \lambda_{\max}(\mathrm{I}_n(\boldsymbol{\beta}_0))n\alpha_n^2\|\mathbf{u}\|^2 \\
&\leqslant M_2 n p_n \alpha_n^2 \|\mathbf{u}\|^2.
\end{aligned}$$

Therefore, $|I_{n11}| = O_P(\alpha_n\sqrt{np_n})\|\mathbf{u}\| = O_P(n\alpha_n^2)\|\mathbf{u}\|$. For the second term I_{n12}, we have

$$\begin{aligned} I_{n12} &= \frac{1}{2}\alpha_n^2\mathbf{u}^\top\{\nabla^2\ell_n(\boldsymbol{\beta}_0) - E_0[\nabla^2\ell_n(\boldsymbol{\beta}_0)]\}\mathbf{u} + \frac{1}{2}\alpha_n^2\mathbf{u}^\top E_0[\nabla^2\ell_n(\boldsymbol{\beta}_0)]\mathbf{u} \\ &= \frac{1}{2}n\alpha_n^2\mathbf{u}^\top\left[\frac{1}{n}\nabla^2\ell_n(\boldsymbol{\beta}_0) + \mathrm{I}_n(\boldsymbol{\beta}_0)\right]\mathbf{u} - \frac{1}{2}n\alpha_n^2\mathbf{u}^\top\mathrm{I}_n(\boldsymbol{\beta}_0)\mathbf{u} \\ &\triangleq I_{n121} + I_{n122}. \end{aligned}$$

According to Lemma A.1,

$$\begin{aligned} |I_{n121}| &\leqslant \frac{1}{2}n\alpha_n^2\|\mathbf{u}\|^2\left\|\frac{1}{n}\nabla^2\ell_n(\boldsymbol{\beta}_0) + \mathrm{I}_n(\boldsymbol{\beta}_0)\right\| \\ &= \frac{1}{2}n\alpha_n^2\|\mathbf{u}\|^2 o_P\left(\frac{1}{\sqrt{p_n}}\right) = n\alpha_n^2\|\mathbf{u}\|^2 o_P(1). \end{aligned}$$

For the third term I_{n13}, let

$$A^{(n)} = \frac{1}{n}\sum_{i=1}^{n} H(\mathbf{W}_i)\mathbf{X}_i\mathbf{X}_i^\top - E_0[H(\mathbf{W}_1)\mathbf{X}_1\mathbf{X}_1^\top].$$

Under conditions (A4) and (A5), we have $\|A^{(n)}\| = o_P(1)$, since for every $\eta > 0$,

$$\begin{aligned} \Pr(\|A^{(n)}\| \geqslant \eta) &\leqslant \frac{E_0\|A^{(n)}\|^2}{\eta^2} \\ &= \frac{1}{\eta^2 n^2}\sum_{j,k=1}^{p_n}\sum_{i=1}^{n} E_0\Big\{H(\mathbf{W}_i)X_{ij}X_{ik} - E_0[H(\mathbf{W}_i)X_{ij}X_{ik}]\Big\}^2 \\ &\leqslant \frac{1}{\eta^2 n}\sum_{j,k=1}^{p_n} E_0[H(\mathbf{W}_1)X_{1j}X_{1k}]^2 \\ &\leqslant \frac{1}{\eta^2 n}p_n^2 M_3^2 M_6 \to 0. \end{aligned}$$

So, under condition (A6), we have

$$\begin{aligned}
E_0(I_{n13}^2) &= \frac{1}{36}\alpha_n^6 E_0\left\{\sum_{l=1}^{p_n}\left[\sum_{j,k=1}^{p_n}\sum_{i=1}^{n} d_{n2}(\beta_n^*, \mathbf{W}_i)X_{ij}X_{ik}X_{il}u_j u_k\right]u_l\right\}^2 \\
&\leqslant \frac{1}{36}\alpha_n^6\|\mathbf{u}\|^2 E_0\left[\sum_{i=1}^{n}\left(\sum_{l=1}^{p_n}X_{il}\right)d_{n2}(\beta_n^*, \mathbf{W}_i)\mathbf{u}^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{u}\right]^2 \\
&\leqslant \frac{1}{36}\alpha_n^6\|\mathbf{u}\|^2 E_0\left[\sum_{i=1}^{n}\left|\sum_{l=1}^{p_n}X_{il}\right|\cdot|d_{n2}(\beta_n^*, \mathbf{W}_i)|\mathbf{u}^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{u}\right]^2 \\
&\leqslant \frac{1}{36}\alpha_n^6 p_n^2 M_3^2\|\mathbf{u}\|^2 E_0\left[\mathbf{u}^\top\sum_{i=1}^{n}H(\mathbf{W}_i)\mathbf{X}_i\mathbf{X}_i^\top\mathbf{u}\right]^2 \\
&\leqslant \frac{1}{18}n^2\alpha_n^6 p_n^2 M_3^2\|\mathbf{u}\|^2\left\{E_0(\mathbf{u}^\top A^{(n)}\mathbf{u})^2 + E_0[\mathbf{u}^\top E_0(H(\mathbf{W}_1)\mathbf{X}_1\mathbf{X}_1^\top)\mathbf{u}]^2\right\} \\
&= \frac{1}{18}n^2\alpha_n^6 p_n^2 M_3^2\|\mathbf{u}\|^2\left\{E_0(\mathbf{u}^\top A^{(n)}\mathbf{u})^2 + E_0[\mathbf{u}^\top E_0(H(\mathbf{W}_1)\mathbf{X}_1\mathbf{X}_1^\top)\mathbf{u}]^2\right\} \\
&\leqslant \frac{1}{18}n^2\alpha_n^6 p_n^2 M_3^2\|\mathbf{u}\|^2\left\{E_0\left[E_0\left((\mathbf{u}^\top A^{(n)}\mathbf{u})^2\middle|\|A^{(n)}\|\leqslant\frac{\rho_{n2}}{2}\right)\right] + \rho_{n2}^2\|\mathbf{u}\|^4\right\} \\
&\leqslant \frac{1}{18}n^2\alpha_n^6 p_n^2 M_3^2\|\mathbf{u}\|^2\left(\frac{\rho_{n2}^2}{4}\|\mathbf{u}\|^4 + \rho_{n2}^2\|\mathbf{u}\|^4\right) \\
&= O(\alpha_n^6 n^2 p_n^2\|\mathbf{u}\|^4).
\end{aligned}$$

Therefore, $|I_{n13}| = O_P(\alpha_n^3 n p_n\|\mathbf{u}\|^2) = o_P(n\alpha_n^2)\|\mathbf{u}\|^2$. Now we consider the term I_{n2}, by the definition of the SCAD penalty and $n\lambda_n^2 = O(1)$, we can obtain

$$I_{n2} \leqslant nk_n(a+1)\lambda_n^2/2 = O(n\alpha_n^2).$$

Hence, by choosing a sufficient large constant C, all terms are dominated by I_{n122}, which is negative. This completes the proof of the theorem. □

To facilitate the proof of Theorem 3.2, we give the following lemma, which shows that under certain regularity conditions, with proper choice of the tuning parameter, the estimator possesses the sparsity property; that is, the insignificant variables can exactly be estimated by zero with probability tending to 1.

Lemma A.2 (Sparsity). Suppose conditions (A1)–(A6) hold. If $\sqrt{p_n/n}/\lambda_n \to 0$, then with probability tending to 1, for any given $\boldsymbol{\beta}_{1n}$ satisfying $\|\boldsymbol{\beta}_{1n} - \boldsymbol{\beta}_{10}\| = O_P(\sqrt{p_n/n}\,)$ and

any constant C, we have

$$Q_n\left((\boldsymbol{\beta}_{1n}^\top, \mathbf{0}^\top)^\top\right) = \max_{\|\boldsymbol{\beta}_{2n}\| \leqslant C(p_n/n)^{1/2}} Q_n\left((\boldsymbol{\beta}_{1n}^\top, \boldsymbol{\beta}_{2n}^\top)^\top\right).$$

Namely, for the local maximizer $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}^\top, \hat{\boldsymbol{\beta}}_{2n}^\top)^\top$ in Theorem 3.1, we have

$$\Pr(\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}) \to 1. \qquad \square$$

Proof. Let $\epsilon_n = C\sqrt{p_n/n}$. It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_{1n}$ satisfying $\|\boldsymbol{\beta}_{1n} - \boldsymbol{\beta}_{10}\| = O_P(\sqrt{p_n/n}\,)$, we have

$$\frac{\partial Q_n(\boldsymbol{\beta}_n)}{\partial \beta_{nj}} < 0, \quad \text{if} \quad 0 < \beta_{nj} < \epsilon_n,$$

$$\frac{\partial Q_n(\boldsymbol{\beta}_n)}{\partial \beta_{nj}} > 0, \quad \text{if} \quad -\epsilon_n < \beta_{nj} < 0.$$

where $j = k_n + 1, \ldots, p_n$.

Since $\sqrt{p_n/n}/\lambda_n \to 0$ and $\|\boldsymbol{\beta}_{2n}\| \leqslant C\sqrt{p_n/n}$, by the Taylor expansion we have

$$\begin{aligned}
\frac{\partial Q_n(\boldsymbol{\beta}_n)}{\partial \beta_{nj}} &= \frac{\partial \ell_n(\boldsymbol{\beta}_n)}{\partial \beta_{nj}} - n\lambda_n \mathrm{sgn}(\beta_{nj}) \\
&= \frac{\partial \ell_n(\boldsymbol{\beta}_0)}{\partial \beta_{nj}} + \sum_{k=1}^{p_n} \frac{\partial^2 \ell_n(\boldsymbol{\beta}_0)}{\partial \beta_{nj}\partial \beta_{nk}}(\beta_{nk} - \beta_{0k}) \\
&\quad + \frac{1}{2}\sum_{k,l=1}^{p_n} \frac{\partial^3 \ell_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}\partial \beta_{nk}\partial \beta_{nl}}(\beta_{nk} - \beta_{0k})(\beta_{nl} - \beta_{0l}) - n\lambda_n \mathrm{sgn}(\beta_{nj}) \\
&\triangleq J_{n1} + J_{n2} + J_{n3} + J_{n4},
\end{aligned}$$

where $\boldsymbol{\beta}_n^*$ is a vector between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_0$, and $\mathrm{sgn}(t) = -1, 0$ or 1 if $t < 0, = 0$ or > 0. Now we consider the first term J_{n1}. By conditions (A1) and (A4), we have

$$\begin{aligned}
E_0(J_{n1}^2) &= E_0\left[\sum_{i=1}^n D(\xi(\boldsymbol{\beta}_0, \mathbf{W}_i)) g'(\boldsymbol{\beta}_0^\top \mathbf{X}_i) X_{ij}\right]^2 \\
&= nE_0\left[D^2(\xi(\boldsymbol{\beta}_0, \mathbf{W}_1))(g'(\boldsymbol{\beta}_0^\top \mathbf{X}_1))^2 X_{1j}^2\right] \\
&\leqslant nM_3^2 M_5^{1/2},
\end{aligned}$$

so that

$$J_{n1} = O_P(\sqrt{n}\,) = o_P(\sqrt{np_n}\,). \tag{A.2}$$

For the second term J_{n2},

$$\begin{aligned} J_{n2} &= \sum_{k=1}^{p_n} \left[\frac{\partial^2 \ell_n(\boldsymbol{\beta}_0)}{\partial\beta_{nj}\partial\beta_{nk}} + \mathrm{I}_{njk}(\boldsymbol{\beta}_0) \right] (\beta_{nk} - \beta_{0k}) - \sum_{k=1}^{p_n} \mathrm{I}_{njk}(\boldsymbol{\beta}_0)(\beta_{nk} - \beta_{0k}) \\ &\triangleq J_{n11} + J_{n12}, \end{aligned}$$

where $\mathrm{I}_{njk}(\boldsymbol{\beta}_0)$ is the (j,k)-th cell element of $\mathrm{I}_n(\boldsymbol{\beta}_0)$,

$$\begin{aligned} |J_{n11}| &\leqslant \left\{ \sum_{k=1}^{p_n} \left[\frac{\partial^2 \ell_n(\boldsymbol{\beta}_0)}{\partial\beta_{nj}\partial\beta_{nk}} + \mathrm{I}_{njk}(\boldsymbol{\beta}_0) \right]^2 \right\}^{\frac{1}{2}} \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\| \\ &= O_P(\sqrt{np_n}\,) O_P\left(\sqrt{\frac{p_n}{n}} \right) = o_P(\sqrt{np_n}\,) \end{aligned}$$

and

$$\begin{aligned} |J_{n12}| &\leqslant n \left\{ \sum_{k=1}^{p_n} \mathrm{I}_{njk}^2(\boldsymbol{\beta}_0) \right\}^{\frac{1}{2}} \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\| \\ &\leqslant n\lambda_{\max}\{\mathrm{I}_n(\boldsymbol{\beta}_0)\} O_P\left(\sqrt{\frac{p_n}{n}} \right) = O_P(\sqrt{np_n}\,). \end{aligned}$$

So we have

$$|J_{n2}| = O_P(\sqrt{np_n}\,). \tag{A.3}$$

For the third term J_{n3}, by the condition (A6), we obtain

$$\begin{aligned}
E_0(J_{n3}^2) &= \frac{1}{4}E_0\left\{\sum_{k,l=1}^{p_n}\left[\sum_{i=1}^{n} d_{n2}(\boldsymbol{\beta}_n^*, \mathbf{W}_i)X_{ij}X_{ik}X_{il}\right](\beta_{nk}-\beta_{0k})(\beta_{nl}-\beta_{0l})\right\}^2 \\
&= \frac{1}{4}E_0\left\{\sum_{i=1}^{n}\left[\sum_{k,l=1}^{p_n} X_{ik}X_{il}(\beta_{nk}-\beta_{0k})(\beta_{nl}-\beta_{0l})\right] d_{n2}(\boldsymbol{\beta}_n^*, \mathbf{W}_i)X_{ij}\right\}^2 \\
&\leqslant \frac{1}{4}E_0\left\{\sum_{i=1}^{n}\left[(\beta_n-\beta_0)^\top \mathbf{X}_i\mathbf{X}_i^\top(\beta_n-\beta_0)\right]|d_{n2}(\boldsymbol{\beta}_n^*, \mathbf{W}_i)|\cdot|X_{ij}|\right\}^2 \\
&\leqslant \frac{1}{4}M_3^2E_0\left[(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^\top\sum_{i=1}^{n}H(\mathbf{W}_i)\mathbf{X}_i\mathbf{X}_i^\top(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)\right]^2 \\
&\leqslant \frac{1}{2}n^2M_3^2E_0\left\{E_0\left[\left((\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^\top A^{(n)}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)\right)^2\middle|\|A^{(n)}\|\leqslant\frac{\rho_{n2}}{2}\right]\right\} \\
&\quad +\frac{1}{2}n^2M_3^2\rho_{n2}^2O\left(\frac{p_n^2}{n^2}\right) \\
&\leqslant \frac{1}{2}n^2M_3^2\frac{\rho_{n2}^2}{4}O\left(\frac{p_n^2}{n^2}\right)+\frac{1}{2}n^2M_3^2\rho_{n2}^2O\left(\frac{p_n^2}{n^2}\right)=O(p_n^2).
\end{aligned}$$

Hence

$$|J_{n3}| = O_P(p_n) = o_P(\sqrt{np_n}\,). \tag{A.4}$$

From (A.2)–(A.4), we have

$$\begin{aligned}
\frac{\partial Q_n(\boldsymbol{\beta}_n)}{\partial\beta_{nj}} &= O_P(\sqrt{np_n}\,) - n\lambda_n\mathrm{sgn}(\beta_{nj}) \\
&= n\lambda_n\left[O_P\left(\sqrt{p_n/n}/\lambda_n\right)-\mathrm{sgn}(\beta_{nj})\right].
\end{aligned}$$

Since $\sqrt{p_n/n}/\lambda_n \to 0$, it is clear that the sign of $\partial Q_n(\boldsymbol{\beta}_n)/\partial\beta_{nj}$ is completely determined by the sign of β_{nj}. Therefore, Lemma A.2 follows. □

Proof of Theorem 3.2. As shown in Theorem 3.1, there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ of $Q_n(\boldsymbol{\beta}_n)$. It follows from Lemma A.2 that part (1) holds. Now we prove part (2). From Theorem 3.1, we obtain $\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\| = O_P(\sqrt{p_n/n}\,)$. Using the condition $\min_{1\leqslant j\leqslant k_n}|\beta_{0j}|/\lambda_n\to\infty$, with probability tending to 1, all $\hat{\beta}_{nj}$ $(j=1,\ldots,k_n)$ are bounded away from $[-a\lambda_n, a\lambda_n]$.

In addition, $\Pr(\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}) \to 1$. Thus, with probability tending to 1, we have

$$\nabla Q_n\left((\hat{\boldsymbol{\beta}}_{1n}^\top, \mathbf{0}^\top)^\top\right) = \frac{\partial \ell_n\left((\hat{\boldsymbol{\beta}}_{1n}^\top, \mathbf{0}^\top)^\top\right)}{\partial \boldsymbol{\beta}_{1n}} = \mathbf{0}.$$

For simplicity, let $\ell_n\left((\hat{\boldsymbol{\beta}}_{1n}^\top, \mathbf{0}^\top)^\top\right) \triangleq \ell_{1n}(\hat{\boldsymbol{\beta}}_{1n})$ and $\mathrm{I}_n\left((\boldsymbol{\beta}_{10}^\top, \mathbf{0}^\top)^\top\right) \triangleq \mathrm{I}_n(\boldsymbol{\beta}_{10})$. Using the Taylor expansion on $\partial \ell_{1n}(\hat{\boldsymbol{\beta}}_{1n})/\partial \boldsymbol{\beta}_{1n}$ around $\boldsymbol{\beta}_{10}$, we have

$$\begin{aligned}
\mathbf{0} &= \frac{\partial \ell_{1n}(\hat{\boldsymbol{\beta}}_{1n})}{\partial \boldsymbol{\beta}_{1n}} \\
&= \frac{\partial \ell_{1n}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}} + \frac{\partial^2 \ell_{1n}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n} \partial \boldsymbol{\beta}_{1n}^\top}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) + \frac{1}{2}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})^\top \nabla^2 \left(\frac{\partial \ell_{1n}(\boldsymbol{\beta}_{1n}^*)}{\partial \boldsymbol{\beta}_{1n}}\right)(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}),
\end{aligned}$$

or,

$$\frac{1}{n}\frac{\partial^2 \ell_{1n}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n} \partial \boldsymbol{\beta}_{1n}^\top}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) = -\frac{1}{n}\frac{\partial \ell_{1n}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}} - \frac{1}{2n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})^\top \nabla^2 \left(\frac{\partial \ell_{1n}(\boldsymbol{\beta}_{1n}^*)}{\partial \boldsymbol{\beta}_{1n}}\right)(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}).$$

Since

$$\begin{aligned}
\left|\left[\frac{1}{n}\frac{\partial^2 \ell_{1n}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n} \partial \boldsymbol{\beta}_{1n}^\top} + \mathrm{I}_n(\boldsymbol{\beta}_{10})\right](\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})\right| &\leqslant \left\|\frac{1}{n}\frac{\partial^2 \ell_{1n}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n} \partial \boldsymbol{\beta}_{1n}^\top} + \mathrm{I}_n(\boldsymbol{\beta}_{10})\right\| \cdot \|\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}\| \\
&= o_P\left(\frac{1}{\sqrt{p_n}}\right) O_P\left(\sqrt{\frac{p_n}{n}}\right) = o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

and

$$\begin{aligned}
& E_0 \left\|\frac{1}{2n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})^\top \nabla^2 \left(\frac{\partial \ell_{1n}(\boldsymbol{\beta}_{1n}^*)}{\partial \boldsymbol{\beta}_{1n}}\right)(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})\right\|^2 \\
= & \frac{1}{4} E_0 \left\{\sum_{l=1}^{p_n}\left[\frac{1}{n}\sum_{j,k=1}^{p_n}\sum_{i=1}^{n} d_{n2}(\boldsymbol{\beta}_{1n}^*, \mathbf{W}_{1i}) X_{1ij} X_{1ik} X_{1il}(\hat{\beta}_{1nj} - \beta_{10j})(\hat{\beta}_{1nk} - \beta_{10k})\right]^2\right\} \\
= & \frac{1}{4} E_0 \left\{\sum_{l=1}^{p_n}\left[\frac{1}{n}\sum_{i=1}^{n} d_{n2}(\boldsymbol{\beta}_{1n}^*, \mathbf{W}_{1i}) X_{1il}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})^\top \mathbf{X}_{1i} \mathbf{X}_{1i}^\top (\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})\right]^2\right\} \\
\leqslant & \frac{1}{4} E_0 \left\{\sum_{l=1}^{p_n}\left[\frac{1}{n}\sum_{i=1}^{n} |d_{n2}(\boldsymbol{\beta}_{1n}^*, \mathbf{W}_{1i})| \cdot |X_{1il}|(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})^\top \mathbf{X}_{1i} \mathbf{X}_{1i}^\top (\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})\right]^2\right\} \\
\leqslant & \frac{1}{4} M_3^2 p_n E_0 \left[\frac{1}{n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})^\top \sum_{i=1}^{n} H(\mathbf{W}_{1i}) \mathbf{X}_{1i} \mathbf{X}_{1i}^\top (\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})\right]^2 \\
= & o\left(p_n \frac{p_n^2}{n^2}\right) = o\left(\frac{1}{n}\right).
\end{aligned}$$

Therefore,

$$\mathrm{I}_n(\boldsymbol{\beta}_{10})(\hat{\boldsymbol{\beta}}_{1n}-\boldsymbol{\beta}_{10})=\frac{1}{n}\frac{\partial \ell_{1n}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}}+o_P\left(\frac{1}{\sqrt{n}}\right).$$

For any $k_n \times 1$ scalar vector $\boldsymbol{\alpha}$, we have

$$\begin{aligned}\sqrt{n}\boldsymbol{\alpha}^\top \mathrm{I}_n^{\frac{1}{2}}(\boldsymbol{\beta}_{10})(\hat{\boldsymbol{\beta}}_{1n}-\boldsymbol{\beta}_{10}) &= \frac{1}{\sqrt{n}}\boldsymbol{\alpha}^\top \mathrm{I}_n^{-\frac{1}{2}}(\boldsymbol{\beta}_{10})\frac{\partial \ell_{1n}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}}+o_P(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\alpha}^\top \mathrm{I}_n^{-\frac{1}{2}}(\boldsymbol{\beta}_{10})\frac{\partial \ell_{1ni}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}}+o_P(1).\end{aligned}$$

Let

$$V_{ni}=\frac{1}{\sqrt{n}}\boldsymbol{\alpha}^\top \mathrm{I}_n^{-\frac{1}{2}}(\boldsymbol{\beta}_{10})\frac{\partial \ell_{1ni}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}},$$

then $E_0(V_{ni})=0$ and

$$E_0(V_{ni}^2)=\frac{1}{n}\boldsymbol{\alpha}^\top \mathrm{I}_n^{-\frac{1}{2}}(\boldsymbol{\beta}_{10})E_0\left\{\left[\frac{\partial \ell_{1ni}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}}\right]\left[\frac{\partial \ell_{1ni}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}}\right]^\top\right\}\mathrm{I}_n^{-\frac{1}{2}}(\boldsymbol{\beta}_{10})\boldsymbol{\alpha}=\frac{1}{n}.$$

We only need to verify the condition of the Lindeberg–Feller central limit theorem. Namely, for every $\epsilon>0$,

$$\lim_{n\to\infty}\sum_{i=1}^{n}E_0\left[V_{ni}^2 I(|V_{ni}|\geqslant\epsilon)\right]=0.$$

By the Hölder inequality, we have

$$\begin{aligned}\sum_{i=1}^{n}E_0\left[V_{ni}^2 I(|V_{ni}|\geqslant\epsilon)\right] &= nE_0\left[V_{n1}^2 I(|V_{n1}|\geqslant\epsilon)\right]\\ &\leqslant n[E_0(V_{n1}^4)]^{\frac{1}{2}}\cdot[\Pr(|V_{n1}|\geqslant\epsilon)]^{\frac{1}{2}}.\end{aligned}$$

Under conditions (A3) and (A4), we obtain

$$\Pr(|V_{n1}|\geqslant\epsilon)\leqslant\frac{E_0(V_{n1}^2)}{\epsilon^2}=O\left(\frac{1}{n}\right)$$

and

$$\begin{aligned} E_0(V_{n1}^4) &= \frac{1}{n^2} E_0 \left\{ \left[\frac{\partial \ell_{1n1}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}} \right]^\top \mathrm{I}_n^{-\frac{1}{2}}(\boldsymbol{\beta}_{10}) \boldsymbol{\alpha}\boldsymbol{\alpha}^\top \mathrm{I}_n^{-\frac{1}{2}}(\boldsymbol{\beta}_{10}) \left[\frac{\partial \ell_{1n1}(\boldsymbol{\beta}_{10})}{\partial \boldsymbol{\beta}_{1n}} \right] \right\}^2 \\ &\leqslant \frac{1}{n^2} \lambda_{\max}(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top) \lambda_{\max}\{\mathrm{I}_n^{-1}(\boldsymbol{\beta}_{10})\} E_0 \left\{ \sum_{j=1}^{k_n} \left[\frac{\partial \ell_{1n1}(\boldsymbol{\beta}_{10})}{\partial \beta_{1nj}} \right]^2 \right\}^2 \\ &\leqslant \frac{1}{n^2} \lambda_{\min}^{-1}\{\mathrm{I}_n(\boldsymbol{\beta}_{10})\} k_n \sum_{j=1}^{k_n} E_0 \left[\frac{\partial \ell_{1n1}(\boldsymbol{\beta}_{10})}{\partial \beta_{1nj}} \right]^4 \\ &= O\left(\frac{p_n^2}{n^2} \right). \end{aligned}$$

Thus we have

$$\sum_{i=1}^{n} E_0 \left[V_{ni}^2 I(|V_{ni}| \geqslant \epsilon) \right] \leqslant O\left(n \frac{p_n}{n} \frac{1}{\sqrt{n}} \right) = o(1).$$

Hence, by the Lindeberg–Feller central limit theorem and Slutsky's theorem, Theorem 3.2 (2) follows. □

Acknowledgments

The authors would like to thank the Editor and referees for their comments and valuable suggestions. Guo-Liang Tian's research was partially supported by a grant (HKU 779210M) from the Research Grant Council of the Hong Kong Special Administrative Region and a grant (Project Code: 2010-1115-9010) from HKU Seed Funding Program for Basic Research. Mingqiu Wang's research was supported by the Foundation of Qufu Normal University (xkj201304) and the NSFC grant (11101063). Lixin Song's research was supported by Specialized Research Fund for the Doctoral Program of Higher Education (20100041110036) and the NSFC grant (61175041).

References

[1] Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. Roy. Statist. Soc. Ser. B* **73**, 325–349.

[2] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351.

[3] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

[4] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70**, 849–911.

[5] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928–961.

[6] Fleming, T. R. and Harrington, D.P . (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[7] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.

[8] Gart, J. J., Krewski, D., Lee, P. N., Tarone, R. E. and Wahrendorf, J. (1986). *Statistical Methods in Cancer Research, Volume III, The Design and Analysis of Longterm Animal Experiments.* IARC Scientific Publications No. 79. International Agency for Research on Cancer, Lyon.

[9] Grummer-Strawn, L. M. (1993). Regression analysis of current status data: An application to breast feeding. *J. Amer. Statist. Assoc.* **88**, 758–765.

[10] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540–568.

[11] Huang, J., Horowitz, J. L. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high–dimensional regression medels. *Ann. Statist.* **36**, 587–613.

[12] Huang, J. and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analy.* **16**, 176–195.

[13] Huang, J., Ma, S. and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62**, 813–820.

[14] Jewell, N. P. and Shiboski, S. C. (1990). Statistical analysis of HIV infectivity based on a partner study. *Biometrics* **46**, 1133–1150.

[15] Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–1378.

[16] Lam, K. F. and Xue, H. (2005). A semiparametric regression cure model with current status data. *Biometrika* **92**, 573–586.

[17] Lin, X. and Wang, L. (2010). A semiparametric probit model for case 2 interval-censored failure time data. *Statist. Med.* **29**, 972–981.

[18] Ma, S. (2009). Cure model with current status data. *Statist. Sinica* **19**, 233–249.

[19] Ma, S. and Du, P. (2012). Variable selection in partly linear regression model with diverging dimensions for right censored data. *Statist. Sinica*. To appear.

[20] Shows, J., Lu, W. and Zhang, H. (2010). Sparse Estimation and Inference for Censored Median Regression. *J. Statist. Plann. Inference* **140**, 1903–1917.

[21] Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

[22] Tibshirani, R. J. (1997). The lasso method for variable selection in the Cox model. *Statist. in Med.* **16** 385–395.

[23] Wang, H., Li, B. and Leng, C. (2009). Skrinkage tuning parameter selection with a diverging number of parameters. *J. Roy. Statist. Soc. Ser. B* **71**, 671–683.

[24] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

[25] Wang, M., Song, L. and Wang, X. (2010). Bridge estimation for generalized linear models with a diverging number of parameters. *Statist. Probab. Lett.* **80**, 1584–1596.

[26] Wang, M., Song, L. and Wang, X. (2011). Bridge estimators in the partially linear model with high dimensionality. *Commun. Statist.–Theory Meth.* **40**, 4325–4346.

[27] Wang, L. and Lin, X. (2011). A Bayesian approach for analyzing case 2 interval-censored data under the semiparametric proportional odds model. *Statist. Probab. Lett.* **81**, 876–883.

[28] Xue, H., Lam, K. F. and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *J. Amer. Statist. Assoc.* **99**, 346–356.

[29] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

[30] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

Table 1: Simulation results for the linear model $Y = \boldsymbol{\beta}_n^\top \boldsymbol{X} + \varepsilon$

Error	(n, p_n)	Method	Correctly fitted	Overfitted	No. of zeros C	No. of zeros IC	Average MSE
Normal	(100,13)	Lasso	0.042(0.201)	0.918(0.275)	7.582(1.527)	0.042(0.211)	0.640(0.346)
		ALasso	0.122(0.328)	0.800(0.400)	8.572(1.025)	0.086(0.308)	0.377(0.330)
		SCAD	0.610(0.488)	0.262(0.440)	9.584(0.782)	0.128(0.334)	0.381(0.467)
		SCAD*	0.698(0.460)	0.210(0.408)	9.680(0.618)	0.096(0.308)	0.363(0.451)
		Oracle	1.000(0.000)	0.000(0.000)	10.000(0.000)	0.000(0.000)	0.176(0.171)
	(300,19)	Lasso	0.004(0.063)	0.996(0.063)	9.398(1.962)	0.000(0.000)	0.577(0.259)
		ALasso	0.243(0.430)	0.757(0.430)	12.013(0.745)	0.000(0.000)	0.243(0.130)
		SCAD	0.860(0.347)	0.140(0.347)	12.834(0.446)	0.000(0.000)	0.155(0.107)
		SCAD*	0.823(0.382)	0.177(0.382)	12.810(0.426)	0.000(0.000)	0.162(0.114)
		Oracle	1.000(0.000)	0.000(0.000)	13.000(0.000)	0.000(0.000)	0.127(0.071)
	(600,24)	Lasso	0.002(0.045)	0.998(0.045)	10.658(2.861)	0.000(0.000)	0.570(0.192)
		ALasso	0.460(0.501)	0.540(0.501)	14.370(0.661)	0.000(0.000)	0.236(0.105)
		SCAD	0.870(0.337)	0.130(0.337)	14.846(0.437)	0.000(0.000)	0.135(0.076)
		SCAD*	0.840(0.368)	0.160(0.368)	14.820(0.435)	0.000(0.000)	0.127(0.067)
		Oracle	1.000(0.000)	0.000(0.000)	15.000(0.000)	0.000(0.000)	0.119(0.060)
Logistic	(100,13)	Lasso	0.192(0.394)	0.414(0.493)	8.810(1.620)	0.490(0.668)	1.729(0.778)
		ALasso	0.286(0.452)	0.274(0.446)	8.918(1.660)	0.576(0.722)	1.426(0.849)
		SCAD	0.338(0.474)	0.126(0.332)	9.652(0.651)	0.640(0.663)	1.100(0.987)
		SCAD*	0.378(0.485)	0.144(0.351)	9.596(0.655)	0.562(0.644)	1.059(0.971)
		Oracle	1.000(0.000)	0.000(0.000)	10.000(0.000)	0.000(0.000)	0.360(0.374)
	(300,19)	Lasso	0.170(0.376)	0.746(0.436)	10.614(3.255)	0.086(0.288)	1.707(0.600)
		ALAsso	0.492(0.500)	0.424(0.495)	12.484(0.589)	0.084(0.278)	0.911(0.430)
		SCAD	0.678(0.468)	0.162(0.369)	12.742(0.576)	0.170(0.402)	0.436(0.364)
		SCAD*	0.718(0.450)	0.154(0.361)	12.780(0.482)	0.130(0.343)	0.411(0.352)
		Oracle	1.000(0.000)	0.000(0.000)	13.000(0.000)	0.000(0.000)	0.269(0.162)
	(600,24)	Lasso	0.084(0.278)	0.914(0.281)	6.728(5.273)	0.002(0.045)	0.983(0.775)
		ALasso	0.698(0.460)	0.286(0.452)	14.692(0.483)	0.016(0.126)	0.959(0.348)
		SCAD	0.852(0.355)	0.132(0.339)	14.832(0.482)	0.018(0.147)	0.272(0.176)
		SCAD*	0.814(0.389)	0.174(0.379)	14.792(0.479)	0.012(0.109)	0.273(0.180)
		Oracle	1.000(0.000)	0.000(0.000)	15.000(0.000)	0.000(0.000)	0.229(0.112)

Table 2: Simulation results for the model $Y = \exp(\boldsymbol{\beta}_n^\top \boldsymbol{X}) + \varepsilon$

Error	(n, p_n)	Method	Correctly fitted	Overfitted	No. of zeros C	No. of zeros IC	Average MSE
Normal	(100,13)	Lasso	0.156(0.363)	0.690(0.463)	8.354(1.393)	0.174(0.429)	0.154(0.111)
		ALasso	0.448(0.498)	0.382(0.486)	9.324(0.939)	0.188(0.435)	0.113(0.100)
		SCAD	0.274(0.446)	0.572(0.495)	8.882(1.148)	0.176(0.435)	0.136(0.133)
		Oracle	1.000(0.000)	0.000(0.000)	10.000(0.000)	0.000(0.000)	0.040(0.038)
	(300,19)	Lasso	0.010(0.100)	0.990(0.100)	9.904(1.797)	0.000(0.000)	0.080(0.042)
		ALasso	0.508(0.500)	0.492(0.500)	12.294(0.837)	0.000(0.000)	0.040(0.023)
		SCAD	0.478(0.500)	0.522(0.500)	12.258(0.939)	0.000(0.000)	0.040(0.030)
		Oracle	1.000(0.000)	0.000(0.000)	13.000(0.000)	0.000(0.000)	0.026(0.017)
	(600,24)	Lasso	0.000(0.000)	1.000(0.000)	10.532(2.111)	0.000(0.000)	0.067(0.032)
		ALasso	0.448(0.498)	0.552(0.498)	14.324(0.699)	0.000(0.000)	0.026(0.014)
		SCAD	0.640(0.480)	0.360(0.480)	14.590(0.622)	0.000(0.000)	0.022(0.014)
		Oracle	1.000(0.000)	0.000(0.000)	15.000(0.000)	0.000(0.000)	0.017(0.009)
Logistic	(100,13)	Lasso	0.140(0.347)	0.380(0.486)	8.962(1.205)	0.654(0.769)	0.291(0.219)
		ALasso	0.190(0.393)	0.256(0.437)	9.324(1.022)	0.744(0.769)	0.285(0.233)
		SCAD	0.206(0.405)	0.266(0.442)	9.316(1.038)	0.742(0.793)	0.338(0.288)
		Oracle	1.000(0.000)	0.000(0.000)	10.000(0.000)	0.000(0.000)	0.081(0.089)
	(300,19)	Lasso	0.160(0.367)	0.814(0.389)	11.092(1.532)	0.026(0.159)	0.129(0.075)
		ALasso	0.552(0.498)	0.414(0.493)	12.364(0.888)	0.036(0.197)	0.081(0.066)
		SCAD	0.732(0.443)	0.222(0.416)	12.670(0.706)	0.052(0.248)	0.063(0.064)
		Oracle	1.000(0.000)	0.000(0.000)	13.000(0.000)	0.000(0.000)	0.039(0.028)
	(600,24)	Lasso	0.128(0.334)	0.872(0.334)	12.330(2.292)	0.000(0.000)	0.097(0.112)
		ALasso	0.716(0.451)	0.282(0.450)	14.414(1.424)	0.002(0.045)	0.074(0.147)
		SCAD	0.780(0.415)	0.220(0.415)	14.370(1.859)	0.000(0.000)	0.056(0.126)
		Oracle	1.000(0.000)	0.000(0.000)	15.000(0.000)	0.000(0.000)	0.030(0.022)

Table 3: Results for primary biliary cirrhosis data

	MLE		Lasso(T)		Lasso		ALasso		SCAD	
Variables	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
X_1	0.096	0.094	0.00	0.00	0.000	0.000	0.000	0.000	0.000	0.000
X_2	−0.265	0.104	0.17	0.09	−0.203	0.055	−0.228	0.061	−0.339	0.092
X_3	0.065	0.095	−0.01	0.03	0.005	0.004	0.000	0.000	0.000	0.000
X_4	−0.266	0.159	0.04	0.07	−0.149	0.047	−0.142	0.051	0.000	0.000
X_5	0.019	0.110	0.00	0.00	0.000	0.000	0.000	0.000	0.000	0.000
X_6	−0.132	0.102	0.02	0.05	−0.064	0.031	0.000	0.000	0.000	0.000
X_7	−0.281	0.136	0.18	0.11	−0.260	0.062	−0.257	0.069	−0.433	0.112
X_8	−0.842	0.216	0.35	0.12	−0.702	0.099	−1.073	0.143	−1.212	0.162
X_9	−0.022	0.115	0.00	0.01	0.000	0.000	0.000	0.000	0.000	0.000
X_{10}	0.149	0.111	−0.22	0.10	0.113	0.043	0.026	0.011	0.000	0.000
X_{11}	−0.179	0.124	0.21	0.11	−0.223	0.059	−0.136	0.042	−0.021	0.008
X_{12}	−0.234	0.101	0.00	0.00	−0.138	0.048	−0.161	0.052	−0.306	0.095
X_{13}	−0.180	0.104	0.09	0.08	−0.093	0.039	−0.011	0.006	0.000	0.000
X_{14}	−0.096	0.121	0.00	0.00	0.000	0.000	0.000	0.000	0.000	0.000
X_{15}	−0.039	0.101	0.00	0.00	0.000	0.000	0.000	0.000	0.000	0.000
X_{16}	−0.346	0.106	0.09	0.09	−0.247	0.060	−0.268	0.070	−0.175	0.056
X_{17}	−0.257	0.115	0.21	0.09	−0.193	0.055	−0.225	0.062	−0.345	0.095